DQO Companion, Version 1.0
User's Guide

Welcome to the *DQO Companion.* See the ReadMe file for known problems and the disclaimer.

What are DQOs?

Data Quality Objectives (DQOs) are criteria that are a part of the output from EPA's Data Quality Objective process. The process is a set of seven steps that help link decision-making goals to environmental data collection methods to ensure that the data collected meet the needs of the decision maker. (Do you have the right type of data to answer the questions that are being asked? Do you have enough data? Is the data of sufficient quality? …) The DQO Companion software is specifically geared towards evaluating errors in decisions based on 3 summary statistics of fine particulate ($PM_{2.5}$) data. The summary statistics are the 3-year annual average concentration, the average of the 98th annual percentiles from 3 consecutive years, and daily concentrations.

What does the software do?

It plots decision performance curves for a variety of input parameters related to measuring $PM_{2.5}$ as it relates to an action limit. For example, the curves show the probability of observing a 3-year annual average that is greater than 16 $\mu g/m^3$ (the action limit) when, in truth, the 3-year annual average is 20 $\mu g/m^3$. These curves let the user choose acceptable decision errors, balancing these errors against the costs of controlling measurement uncertainty such as data completeness, precision and bias. In some cases bias and even precision can be more tightly controlled with additional calibration checks and maintenance checks, which increase sampling costs.

This final balance, however, is also influenced by natural spatial and temporal conditions that vary considerably across the United States. Costs associated with reducing the influence of these natural conditions would be costs associated with increasing sampling frequency and increasing the number of sites in a particular network.

The original DQOs for $PM_{2.5}$ were chosen on the basis of the most conservative "realistic" combination of the natural and measurement conditions, which may not correspond to the conditions that exist within your particular state, reporting organization, region or MSA. Hence, the software allows the user to input parameters describing the natural variation that may be more realistic for his/her data.

How does it do it?

The software simulates 10,000 3-year scenarios for the parameters that are entered and a range of 3-year means and mean 98th percentiles. It simulates both truth and "biased" measurements. It then looks at the percent of the time that the measured values indicate that the action limits have been violated. These are plotted as decision performance curves.

For a given case, say a 3-year annual mean of 13.2 $\mu g/m^3$ or a mean 98th percentile of 71.3 $\mu g/m^3$, the software effectively simulates 10,000 different sets of three years worth of truth with a mean of 13.2 $\mu g/m^3$ and 10,000 more with a 98th percentile of 71.3 $\mu g/m^3$. Each of these sets of "truth" are based on population parameters that have been supplied by the user, including the degree of seasonality (measured by the "ratio"), the degree of randomness about the seasonal variation (measured by the "population CV"), and effectively a measure of how fast the natural conditions can change from day-to-day (measured by the "autocorrelation"). Next, the data are sampled once every "m" days. Then, for each quarter, the software randomly throws out data to mimic data "completeness." Lastly, a random amount of measurement error is added or subtracted for each day. The measurement biases are always taken to be at the maximum amount allowed, with both positive and negative bias being simulated.

The software produces two curves. The left curve shows the probability (power) of observing a value above the action limit, given that there is a positive bias in the measurement system. The right curve shows the probability of observing a value above the action limit, given that there is a negative bias in the measurement system. About 6 billion values are manipulated in constructing the two curves. Once the curves are constructed, the gray zones are found by estimating where the curves cross the chosen values for the type I and II error bounds.

What do the parameters represent and how do I tell what values to input?

First, get a copy of the G-4 guidance document, http://www.epa.gov/quality/qa_docs.html. It doesn't need to be read in full, but it is handy as a reference. Also, get a copy of the $PM_{2.5}$ DQO document or the model QAPP section 7, http://www.epa.gov/ttn/amtic/files/ambient/pm25/qa/totdoc.pdf.

Second, gather all the information and people that are needed. At least one person should read all of Step One in G-4. In particular, it describes who needs to be a part of the process. The data that you need should be representative of your geographic region. As a start, get all of the $PM_{2.5}$ mass measurements from your region for the past few years from AIRS.

The remainder of this section goes though each of the input parameters, describing what they measure and giving some advice on how to choose appropriate values. It is anticipated that most of the DQO software input values will be calculated by AIRS AQS in the near future.

Parameters that describe the decision errors levels

Type I error

This is the probability of observing a value above the action limit when it truly is below the action limit. The model QAPP and EPA guidance use 5% (entered as 0.05) for the Type I error. This parameter is limited to be greater than 1%, as otherwise more simulations are needed to get robust results. See Step 6 of G-4 for additional guidance.

Type II error

This is the probability of observing a value below the action limit when it truly is above the action limit. Since the curves show the probability of observing a value above the action limit, the value of 1 minus the type II error is shown at the top of the graphs. The model QAPP and EPA guidance use 5% (entered as 0.05) for the Type II error. This parameter is limited to be greater than 1%, as otherwise more simulations are needed to get robust results. See Step 6 of G-4 for additional guidance.

Note the choice of the labels Type I and Type II depend on your point of view. The choice shown by the software and described above is based on what was first programmed. These definitions should not be taken as a requirement.

Parameters that describe the action limit

Annual Average Action Limit

The default action limit is 15.0 μg/m^3. Note that the software adds 0.05 to the user-specified action limit. Also, note that the simulated values are rounded to the nearest tenth of a microgram per cubic meter before comparison to this parameter. Valid values for this action limit are between 5 and 30.

98th Percentile Action Limit

The default action limit is 65 μg/m^3. Note that the simulated values are rounded to the nearest microgram per cubic meter before comparison to this parameter. Valid values for this action limit are between 15 and 100.

Daily Action Limit (under the AQI Tab)

The default action limit is 15.4 μg/m^3. Valid values for this action limit are between 5 and 500.

Parameters that describe the nature of the true PM concentrations

These are generally uncontrollable parameters that have a strong affect on the decision errors. They are connected with the model that is used in the simulations. This section describes both the parameters and some ways of estimating the parameters from data that are available in AIRS. Initial investigations have shown that these parameters vary geographically.

The basic model is that the true PM levels vary about a sinusoidal curve with one full oscillation in each year. Three parameters describe characteristics of the sine curve and the natural deviations from the sine curve.

Ratio (Seasonality)

The ratio parameter is a measure of the degree of seasonality in the data. It is the ratio of the high point to the low point on the sine curve. The model assumes that the amplitude of the sine curve is proportional to the mean. If you have at least a year of data, then the ratio can be estimated by calculating the means for each month and dividing the highest value by the lowest. If you have more than 1 year of data (great!), then for each month average all of the data even though it may come from different years. Even though it is tempting, do not use the ratio of the maximum concentration to the minimum concentration. The maximum and minimum are too variable. Use the ratio of the monthly averages.

Population CV

This parameter measures the amount of random, day-to-day variation of the true concentration about the sine curve. This parameter is a bit harder to estimate. The following does a reasonable job. Start with every 6th day measurements (deleting if needed) and take the natural log of each. Create a new sequence of numbers equal to the differences of successive pairs in the sequence of the logs. Remove every other term in the sequence. Let S = the standard deviation of this set of numbers. An estimate for the population CV is $\sqrt{\left(\exp\left(S^2/2\right)-1\right)}$.

Autocorrelation

The final parameter describing the natural variation of the true concentrations is autocorrelation. This is a measurement of the similarity between successive days. Consider two sets of measurements. First, suppose you had measured the $PM_{2.5}$ concentration on every July 15th for the past five years. You would expect those five values to be rather spread out. The population CV captures how different these measurements are from each other. On the other hand, suppose instead you measure the PM concentration each day from July 15, 2002, to July 20, 2002. These values may not be as spread out as the other set, simply because they are nearer in time to each other. Autocorrelation measures this effect. A good way to think of autocorrelation is it measures how quickly the local concentrations can change. The value of the autocorrelation ranges between 0 and 1. A value of 0 means that the local concentrations change very fast. A value of 1 means that the local concentrations are constant.

Estimating autocorrelation is even harder than estimating the population CV. If you don't have daily measurements to work with, then use 0. Realistically, 0 is the most conservative case and can always be used. Assuming you do have daily measurements, let S6 be the standard deviation computed as in the section on population CV, based on differences of the logs from every 6th day measurements. Let $S1$ be the same thing using differences of logs from every day measurements. If $S6 > S1$ then you have some autocorrelation. You can estimate it with $\left(S6^2 - S1^2\right)/S6^2$. This tends to slightly over estimate the truth. Since it is better to under estimate this parameter (to make the results more conservative) you may want to multiply by 0.85.

Do **not** estimate the autocorrelation with the usual correlation estimate between successive values. This does not work when there is seasonality.

Parameters that describe the nature of the sampling

1 in m day sampling

This is the intended sampling frequency. The value of m must be an integer from 1 to 12. (1, 3, 6, and 12 are the most common values.)

Completeness

This is the minimum acceptable percentage of the data that is intended to be collected. This is to mimic random occurrences of data loss, such as a power outage on a scheduled sampling day. The criterion is applied quarterly. Enter 75% completeness as 0.75.

Parameters that describe the nature of the measurement error

Bias

This is the maximum allowable measurement bias as a fraction of the truth. Bias is a consistent measurement error. Enter a positive value (0.1 for 10%). Both positive and negative biases are simulated. Consult the $PM_{2.5}$ QA Reports for estimates of bias appropriate for various areas and time periods, http://www.epa.gov/ttn/amtic/pmqagen.html.

Measurement CV

This is the random component of the of measurement error expressed as a percent of truth. The random component to the measurement error is assumed to follow a normal distribution with a mean of 0 and a standard deviation that is proportional to truth (for the given day). Enter 0.1 for 10 percent. Consult the PM2.5 QA Reports for estimates of measurement CV (precision) for various areas and time periods, http://www.epa.gov/ttn/amtic/pmqagen.html.

Reporting an AQI

The AQI portion of the software produces decision performance curves for daily measurements. In this case there is less that needs to be considered because there is no need to simulate 3-year periods. (At this time within day variation is not considered, but may at a later point since the AQI is often based on continuous measurements and then aggregated to a daily level.) Hence, the natural variation is "ignored" since truth is considered to consist of a single number.

There are three parameters describing measurement error: bias, drift, and measurement CV. The bias and measurement CV are the same as corresponding parameters for the annual and 98th percentile decisions. The drift parameter is an additional bias term. It has been found that some continuous devices have a "seasonal" bias. So this parameter allows you to separately enter a bound on the bias between your continuous device relative to your FRM (drift) and the bias of the FRM relative to the PEP measurements. Note that drift is measured relative to the "biased" FRM, so if both bias and drift are 10%, then the total bias used is 21 percent. You can leave one or the other of these 0 to select a total bias of your own.